Analysis and Interpretation of Biological Data - Data Analysis Assignment

Report

Submitted by,

Ashutosh Sarda

BE17B012

Implementing Multilayer Perceptron for classifying given MNIST data

The given dataset consists of 800 training data and 200 test data points are run across the neural network for classification into 10 classes (0,9) using multilayer perceptron. Using varying parameters optimal results were obtained and their loss and accuracy were computed.

The model with the best result:

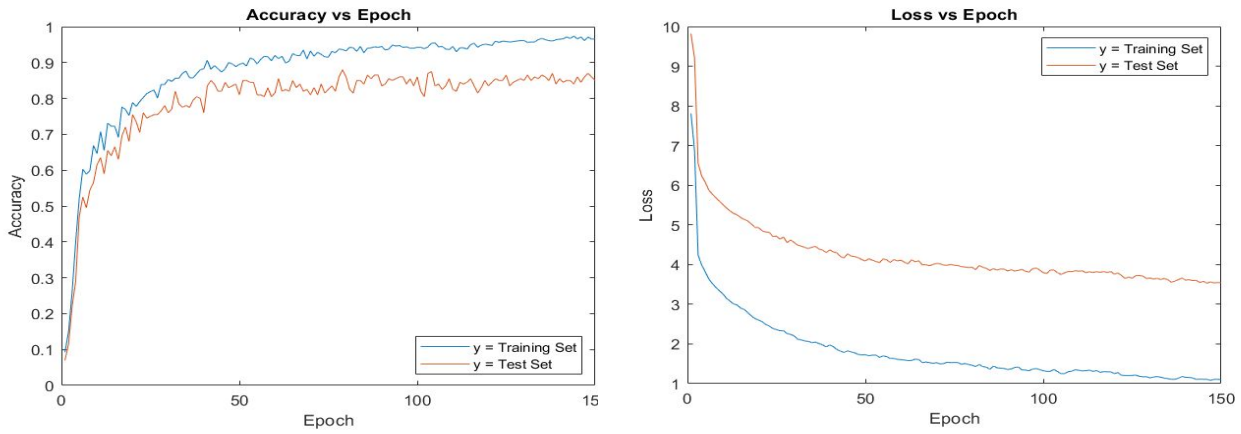

Figure 1: Model with the best result

Model Parameters			
Number of Epochs	150	Regularization Parameter	3
Learning Rate	0.5	Hidden Layer Nodes	100
Training Set Accuracy	95.25%	Test Set Accuracy	86%

Model Specifications:

1. **Network Architecture:** 3 fully connected layers that include the input layer, hidden layer and output layer. The input layer has 784 nodes, the hidden layer has 100 nodes and the output layer has 10 nodes as per the number of classes.

2. **Activation Function:** Sigmoid function has been used as the activation function.

$$g(z) = \frac{1}{1 + e^{-z}}$$

3. **Learning Algorithm:** Stochastic gradient distribution has been used as the learning algorithm.
4. **Learning Rate:** Learning rate of 0.5 was found to be optimal.
5. **Loss for test dataset:** Loss for the test data was found to be 1.098.
6. **Accuracy for the test dataset:** The accuracy was 85% for the test dataset.

Models with varying parameters:

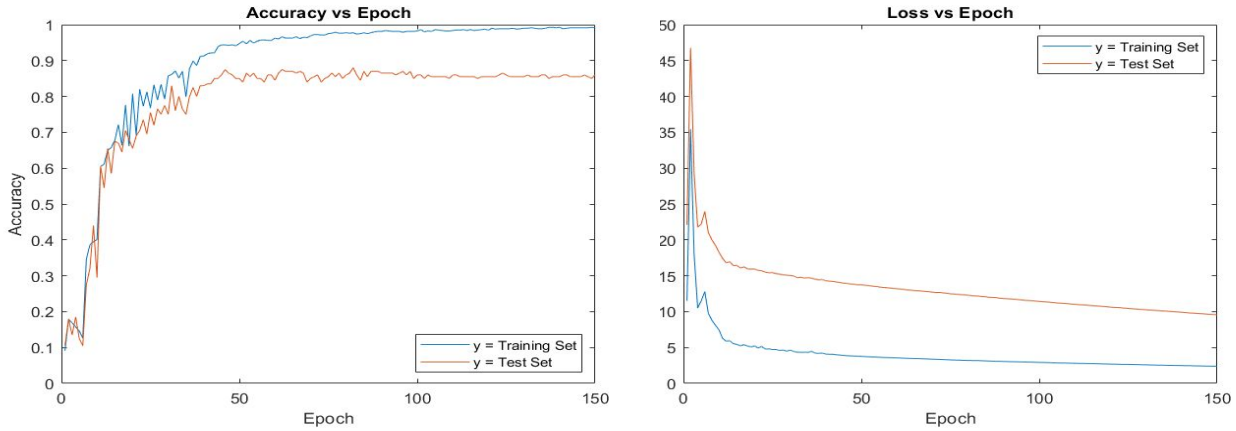

Figure 2: Model with an increased number of nides which results in overfitting.

Model Parameters			
Number of Epochs	150	Regularization Parameter	3
Learning Rate	0.5	Hidden Layer Nodes	100
Training Set Accuracy	99.5%	Test Set Accuracy	84.75%

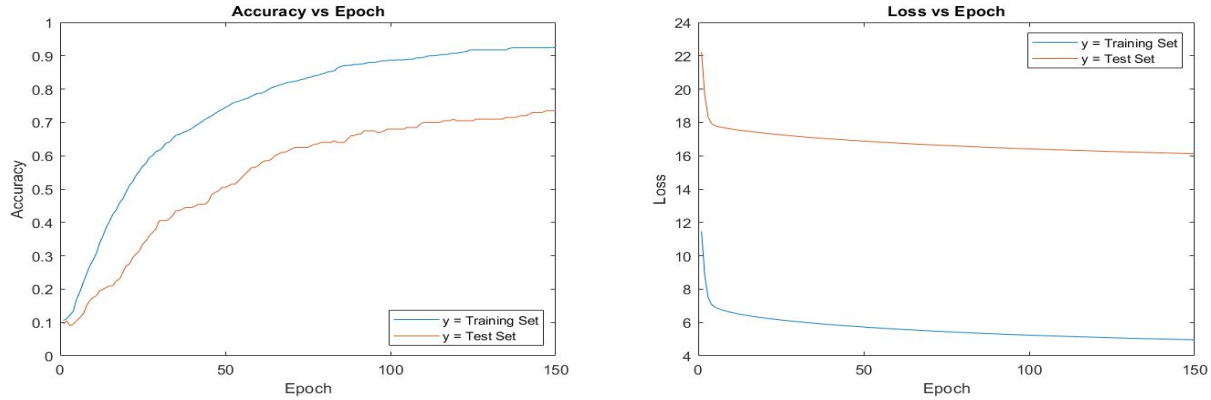

Figure 3: Model with lower learning rate results in lower accuracies for both training and test data. Learning rate hence can be increased.

Model Parameters			
Number of Epochs	150	Regularization Parameter	3
Learning Rate	0.1	Hidden Layer Nodes	100
Training Set Accuracy	90.375%	Test Set Accuracy	73.5%

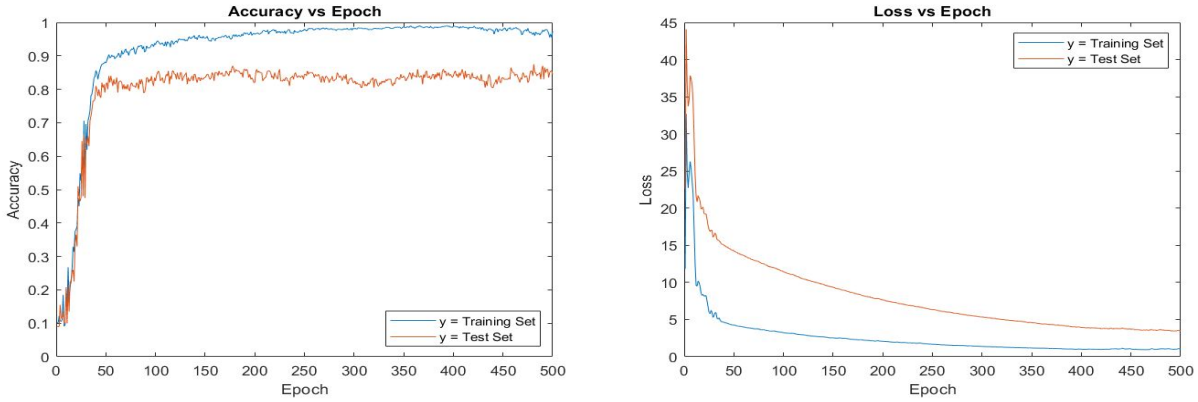

Figure 4: Model with a higher number of epochs results in overfitting as only the training set accuracies have increased.

Model Parameters			
Number of Epochs	500	Regularization Parameter	3
Learning Rate	0.5	Hidden Layer Nodes	100
Training Set Accuracy	98.375%	Test Set Accuracy	83.5%

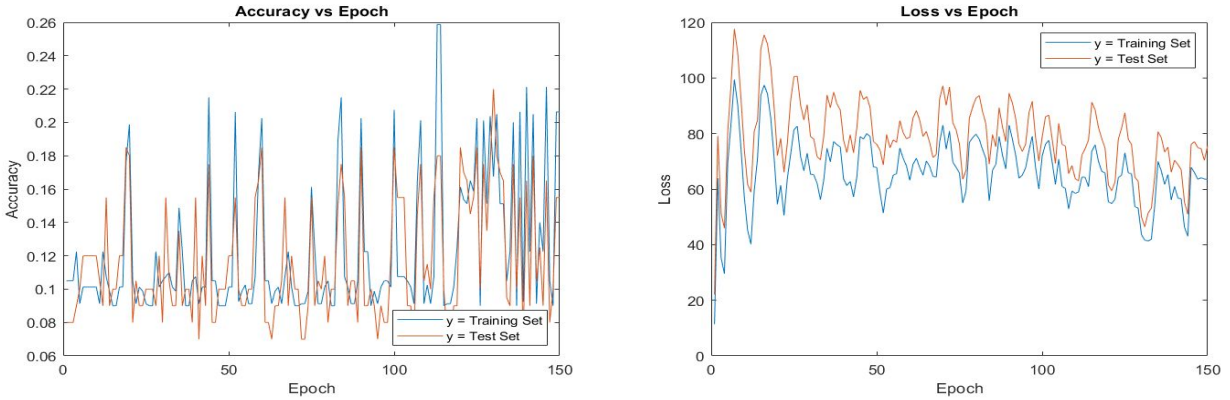

Figure 5: Model with a higher learning rate fails to converge to the minimum loss that results in very poor accuracies for both the datasets.

Model Parameters			
Number of Epochs	150	Regularization Parameter	3
Learning Rate	1	Hidden Layer Nodes	100
Training Set Accuracy	20.625%	Test Set Accuracy	15.5%

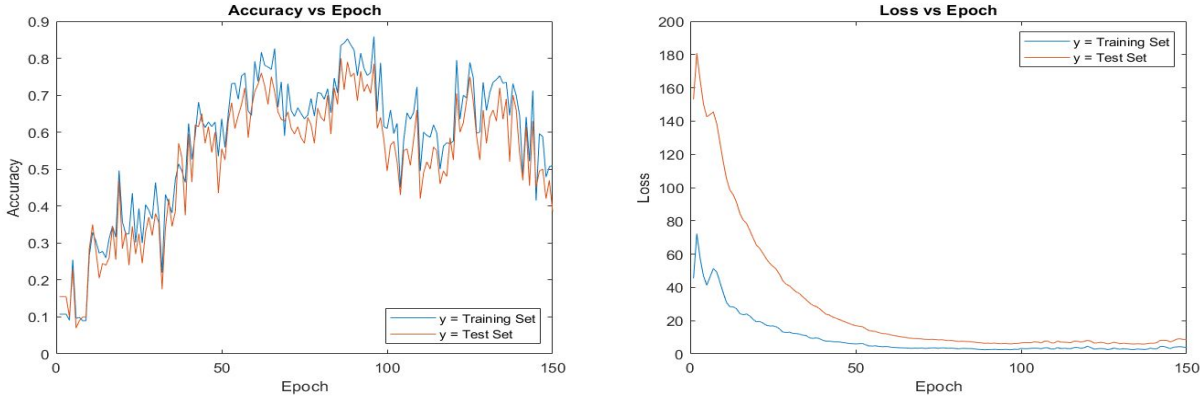

Figure 6: Model with a larger regularization factor (lambda) results in underfitting that gives very poor accuracies for both the datasets.

Model Parameters			
Number of Epochs	150	Regularization Parameter	15
Learning Rate	1	Hidden Layer Nodes	100
Training Set Accuracy	55.125%	Test Set Accuracy	49%

Implementing Density-Based Scan Algorithm to cluster a dataset of 2000 points

The 2000 points dataset were clustered into 4 clusters using a density-based clustering algorithm. Using varying parameters optimum clusters were obtained and observations are shown below.

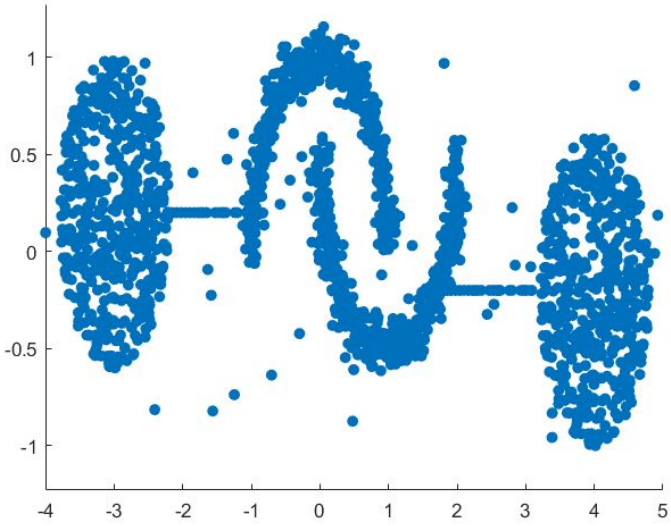

Figure 7: Provided data points

Clusters with optimal parameters:

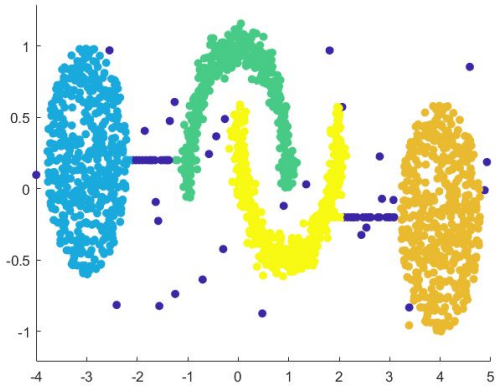

Figure 7: epsilon = 0.17, min points = 23

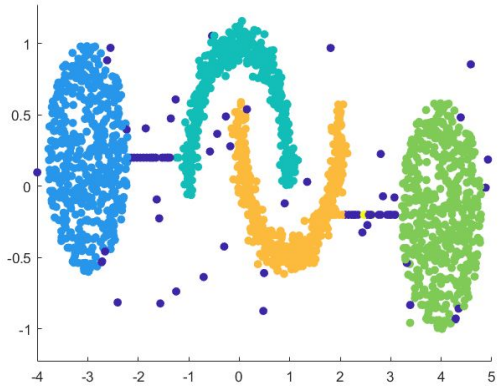

Figure 8: epsilon = 0.2, min points = 15

Clusters with varying parameters:

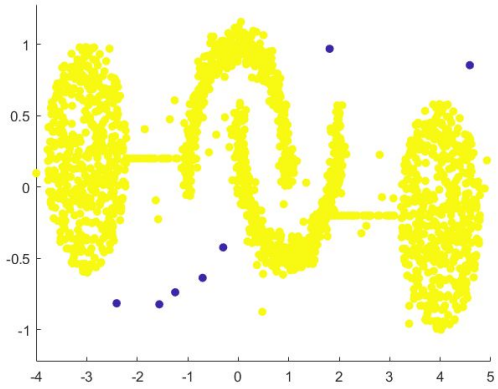

Figure 9: epsilon = 0.5, min points = 15

Increasing the value of epsilon decreases outliers while increasing the cluster size.

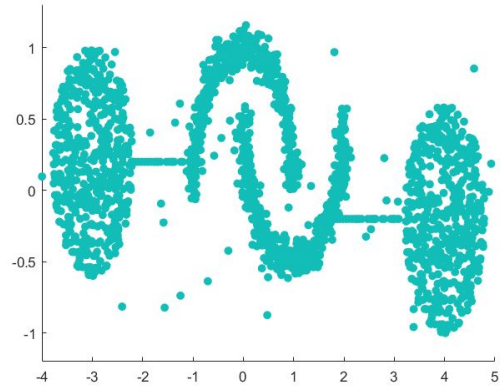

Figure 10: epsilon = 0.05, min points = 15

Decreasing the value of epsilon results in all points being outliers (noise).

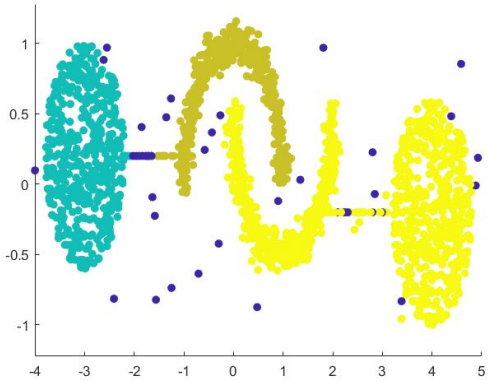

Figure 9: epsilon = 0.5, min points = 15

Intermediate clustering state with intermediate values of parameters.